



夜船有明
YOFUNE LABS

2026 企业 AI Agent 上线前安全检查清单

面向即将接入生产数据的 Copilot、知识库机器人、业务 Agent 与 MCP 工具链。安全部门不必先成为提示词工程师——把一个真实的 Agent 放在桌上，逐项看证据。

- 01 Agent 不是聊天窗口。它拿着业务钥匙：目标、工具、状态。
- 02 没有证据的“是”，不能算通过。账号、工具表、短演示、日志。
- 03 三项阻塞：可以内部试验，不应接生产数据。

本文不是客户交付文件，也不是某次评估报告。实验室路径用于说明机制，不表示已为特定客户完成生产评估。

01 / 如何使用

先选定一个 Agent，再打勾。

不要把这份清单当成问卷发给业务队勾“是”。每一题都要看到证据：账号、工具表、一次短演示、一段日志。演示可以短。没有证据的“我们已经注意了”，不能算通过。十个关键问题里若有三题答成“开发账号共用、工具以后再收、日志还没接好”，这个 Agent 可以继续做内部试验，不应接生产数据。

计分规则

是 + 证据	通过。证据可以是账号截图、工具表、一次演示或一段可回溯日志。
是，但无证据	未通过。口头保证不能写入上线条件。
否 / 以后再收	阻塞项。记录责任人和完成时间，未关闭前不接生产数据。

谁来填

安全负责人和真正挂工具、MCP、账号的人一起填。CISO 可以只看阻塞项和管理层摘要；平台与应用负责人必须能指出具体账号和工具名。抽象的“我们的 AI”无法检查。

本册十节

01	AI Agent 和传统 LLM 应用有什么不同
02	Agent 的资产与信任边界
03	Prompt Injection
04	RAG / Memory
05	Identity & Permission
06	Tool / MCP
07	Sensitive Action
08	Logging & Evidence
09	上线前 30 项检查
10	安全评估应当交付什么



02 / 机制

前八节用来判断机制在不在。勾选之前先读短文：它说明为什么这一项不能只靠提示词。第三节把机制收成三十个可回答是/否的检查。

01 AI Agent 和传统 LLM 应用有什么不同

模型安全要回答：它会不会生成违禁内容、会不会被一套固定句子套出系统提示。这些工作有价值，不够给 Agent 发上线许可。Agent 在模型之外还多了目标、工具和状态。它可以把“帮我处理发票”翻译成一串对内部系统的调用。调用成功时，业务已经改了。这时再讨论模型有没有礼貌地拒答，已经晚了。护栏拦的是话，拦不住工具。

- ☐ 它会在无人确认时调用工具、改数据或发消息，而不只是生成一段文字。
- ☐ 评估标准看行动轨迹（调了什么、以谁的身份、把数据送到哪里），而不是只看模型拒答。
- ☐ 失败之后它会停下来问人，而不是自行换成更强的工具把任务做完。
- ☐ “使用了安全对齐的模型”没有被写成这个 Agent 的风险结论。

失败信号：红队题库全是“请忽略之前的指令”，过关只说明聊天安全。

02 Agent 的资产与信任边界

把模型、Agent、知识库、MCP、账号、外部输入画在一张图上。让未参与搭建的人能看懂数据从哪进来、从哪出去。Memory 要画成独立节点：它同时是数据、策略和攻击入口，一次写入可以在几天后的另一次任务里被执行。能上网的 Agent 和能碰内部名单的 Agent 不应共用一个全能运行时。

- ☐ 模型、Agent、知识库、MCP、账号、外部输入已画在同一张图上。
- ☐ 未参与搭建的人能看懂数据从哪进来、从哪出去。
- ☐ Memory 被画成独立节点，而不是模型旁边的装饰。
- ☐ 能上网或能读附件的 Agent，与能碰内部名单的 Agent 没有共用一个全能运行时。

失败信号：资产清单只有模型供应商名字，没有工具目录和记忆存储。

02 / 机制

03 Prompt Injection

邮件、附件、网页、工单描述进入工作上下文后，作者就获得了对本轮任务的发言权。对模型来说，页面上的隐藏句和用户刚打的那句“帮我看看这份差旅说明”，并没有天然的信任等级。系统提示是建议，不是内核强制。当文档里的指令更具体、更像当前任务时，模型会顺着更具体的那条走。能不能外发，取决于工具是否存在、是否需要审批。

- ☐ 邮件、附件、网页、工单描述、外部 API 已被列为不可信输入。
- ☐ 不可信文本不能直接改写当前目标，也不能直接填进高权限工具参数。
- ☐ 浏览或解析结果带有来源标记，且不能自动变成内部检索词或发送内容。
- ☐ 系统提示中的保密要求从未被当作唯一控制措施。

失败信号：一份版式正常的 PDF 就能把“预审”改写成“外发核对”。

04 RAG / Memory

RAG 的第一道门是哪些文档能进入候选。如果这一步只看相似度、不看身份，后面的提示词和拒答都在补漏。泄露不一定长得像下载：一次正常问答把不该看的段落做成摘要，已经是数据越权。Memory 会把一次注入变成跨会话的规则。写入的时候像偏好，读出的时候像政策。记忆可以提示，不能授权。

- ☐ 检索请求携带调用者身份，并在取片段时过滤；不是全库混检后再让模型“注意保密”。
- ☐ 向量库按租户或部门分库 / 分命名空间。
- ☐ 回答引用的文档编号与权限标签能打开复核。
- ☐ Memory 按用户和任务隔离，有过期、撤销和审计；跨用户复用默认视为越权通道。

失败信号：工艺问题能把质量部带薪酬结构的文件作为“最相关片段”带出来。

02 / 机制

05 Identity & Permission

一次工具调用的日志常常写成：助手访问了订单接口。这句话省略了所有关键主语。发起者、授权范围、执行身份、工具自己的凭证，任何一层被共用，最小权限就只写在纸上。判断权限是否过多，不要看工具名字是否体面，看它能不能完成主任务范围以外的事。能，就过多。提示词里的“请谨慎使用权限”不能单独作为控制措施。

- ☐ 一次工具调用能回溯到具体的人、团队和任务，而不是只看见“助手”。
- ☐ 发起者、授权范围、执行身份、工具自身凭证四层已拆开，没有一把“Agent 专用账号”打通所有工具。
- ☐ 执行身份独立于开发 / 调试账号，且与其他业务队共用。
- ☐ 完成主任务最少需要的权限已写下来；多出来的每一项都有人认领，否则拿掉。

失败信号：同一个运行时服务多个业务队，五分钟内没人能关掉它的写权限。

06 Tool / MCP

MCP 能让内部系统很快被 Agent 用起来，也正因为快，它容易跳过原本针对“新集成”的审批。名字像官方，不表示发布通道受控。来源说不清的 MCP，按未授权软件处理。给任务发牌，不给服务器发万能钥匙。工具表变更视为权限变更。失败后允许自行升级工具，等于把备用路径写成正式路径。

- ☐ 当前任务真正需要的工具已列出，多余的默认拿掉。
- ☐ MCP / 插件的维护方、来源、版本钉死方式和更新审批可以出示。
- ☐ 启动后的实际工具表已存档；增减一项都按权限变更评审。
- ☐ 给 MCP 的凭证按任务裁剪：搜索令牌不具备导出、删除或管理权限。

失败信号：同名 internal-search 在运行时多注册了一个 export_matches。



02 / 机制

07 Sensitive Action

发送、删除、转账、改权限、执行命令，这些能不能在无人确认时发生。一个只负责“查询报销进度”的 Agent，如果同时可以导出全表、发邮件、改供应商信息，它已经不是查询助手。即便从来没人让它改供应商，那把钥匙仍在。高影响工具的接收人、对象、范围应绑定既定目录，不让模型挑选。

- ☐ 发送、删除、转账、改权限、执行命令在无人确认时不能发生。
- ☐ 不可逆动作有独立审批，而不是模型“觉得有必要”即可执行。
- ☐ 主任务范围以外的写操作已经摘除。
- ☐ 记忆或网页内容不能直接填进高权限参数。

失败信号：通知工具的接收人是自由字符串，隐藏段落即可改成外部地址。

08 Logging & Evidence

应用日志记的是 Agent，网关日志记的是服务账号，模型日志记的是会话 ID，工具日志记的是 MCP 名称——四份记录对不上同一个人，就无法判断这是用户授意、Agent 自作主张，还是外部内容改写了目标。出了问题要能切断工具、吊销令牌、冻结记忆，而不是只能关模型。发现的路径必须变成回归项。只修提示词、不留测试，等于没修。

- ☐ 输入、检索片段、工具请求、返回值和人审记录能拼成一次任务的证据链。
- ☐ 应用日志、网关日志、模型会话、MCP 名称能对上同一个人 / 任务。
- ☐ 出了问题能切断工具、吊销令牌、冻结记忆，而不只是关掉模型。
- ☐ 发现的路径已写成可回归的测试项；修复后用同一条路径再走一遍。

失败信号：出事之后只能重装，说不清当时模型看见了哪些工具。

03 / 上线前 30 项

三十项可回答是 / 否的检查

对着一个即将上线的 Agent 逐项勾选。每项只接受“是 + 证据”。把“否”和“以后再收”收成阻塞项。出现三项以上阻塞，明确不得接生产数据。

#	检查项	是	否	证据 (账号 / 日志 / 演示)
01	☐ 工具调用可回溯到人与任务 日志能否从一次调用追到具体的人、团队和任务，而不是只看见“助手”。	☐	☐	
02	☐ 执行身份与开发账号分离 生产任务是否还在使用开发、调试或个人笔记本上的会话。	☐	☐	
03	☐ 四层身份拆开记录 发起者、授权范围、执行身份、工具自身凭证是否分别可查。	☐	☐	
04	☐ 五分钟内可切断写权限 有没有人能在五分钟内关掉写操作、吊销令牌并冻结记忆。	☐	☐	
05	☐ 检索按调用者身份过滤 知识库 / 工单 / 客户库是否按身份取片段，而不是万能服务账号。	☐	☐	
06	☐ 知识库按租户或部门隔离 向量库是否分库或分命名空间，而不是全员混检。	☐	☐	
07	☐ 回答引用可打开复核 用到的文档编号与权限标签能否当场打开。	☐	☐	
08	☐ Memory 按用户和任务隔离 是否禁止跨用户复用，并有过期、撤销和审计。	☐	☐	
09	☐ 外部输入已标记为不可信 邮件、附件、网页、工单、外部 API 是否进入不可信通道。	☐	☐	
10	☐ 不可信文本不能改写目标 外部内容是否无法直接改写当前任务目标。	☐	☐	
11	☐ 浏览结果不能自动变工具参数 抓取或解析文本是否禁止直接填入检索词、收件人或命令。	☐	☐	
12	☐ 提示词保密要求不是唯一控制 是否还有工具裁剪、审批或检索过滤，而不是只靠系统提示。	☐	☐	
13	☐ 当前任务工具表已裁剪 是否只挂主任务需要的工具，多出来的已默认拿掉。	☐	☐	
14	☐ MCP 来源、版本与审批可出示 维护方、镜像或仓库、版本钉死、更新谁批准，能否拿出记录。	☐	☐	
15	☐ 实际工具表已存档并受变更评审 启动后注册出来的函数是否已存档，增减是否当权限变更。	☐	☐	



03 / 上线前 30 项

#	检查项	是	否	证据 (账号 / 日志 / 演示)
16	☐ MCP 凭证按任务最小授权 搜索服务器是否只持有检索令牌，而不持有导出或删除权限。	☐	☐	
17	☐ 失败后禁止自行升级工具 报错后是否必须停下来问人，而不是改调更强接口。	☐	☐	
18	☐ 终端 / 命令执行有沙箱与人审 编码或运维 Agent 是否在隔离工作区，命令是否有允许列表。	☐	☐	
19	☐ 不可逆动作默认需要人确认 发送、删除、转账、改权限、执行命令能否在无人确认时发生。	☐	☐	
20	☐ 发送 / 删除 / 改权限不能由模型自行决定 高影响工具的接收人、对象、范围是否绑定既定目录。	☐	☐	
21	☐ 主任务以外的写操作已摘除 只负责查询的 Agent 是否仍能导出全表、发邮件或改供应商。	☐	☐	
22	☐ 记忆或网页不能填入高权限参数 Memory 与外部正文是否被禁止直接写入特权工具入参。	☐	☐	
23	☐ 一次任务能拼出完整证据链 输入、检索片段、工具请求、返回值、人审是否在一起。	☐	☐	
24	☐ 应用、网关、模型、MCP 日志可对齐 四份记录能否对上同一次任务、同一个人。	☐	☐	
25	☐ 已发现路径写成回归测试 这条路径是否成为固定测试项，而不是口头说过。	☐	☐	
26	☐ 修复后用同一路径复测 已修项是否用同一输入再走一遍，而不是只改了提示词。	☐	☐	
27	☐ 出事有切断工具与冻结记忆的开关 能否在不关闭整个模型服务的情况下停掉行动面。	☐	☐	
28	☐ 资产与信任边界图已画清 模型、Agent、知识库、MCP、账号、外部输入和数据进出是否在一张图上。	☐	☐	
29	☐ 隔离环境已用只读工具跑通 身份、工具表、数据和退出方式是否在写操作之前讲得清。	☐	☐	
30	☐ 三项以上“以后再收”则不接生产数据 共用开发账号、工具以后再收、日志还没接好——出现三项是否明确不得接生产数据。	☐	☐	

04 / 评估交付

一次评估应当留下什么

如果一次所谓评估只留下“整体风险较高”和一页产品截图，管理层无法决策，研发也无法修复。缺下面任何一项，就还不能称为评估完成。KUROSHIO 适合承担规划、探索、工具调用、证据和报告这一段。它不能替代范围确认，也不能替代管理层做上线决定。

- | | |
|--------------------------|---|
| ☐ | 资产与信任边界图
模型、Agent、知识库、MCP、账号、外部输入画在一张图上。未参与搭建的人能看懂数据进出。 |
| ☐ | 风险清单
每条有对象、前提、影响、目前是否可利用。避免把“提示注入”写成没有系统名字的抽象词。 |
| ☐ | 攻击路径
何种输入，经过哪次检索或哪次工具调用，到达何种结果。路径要短，要可复述。 |
| ☐ | 验证证据
输入摘要、关键上下文、工具请求与返回、策略是否拦截。第三人要能核对。 |
| ☐ | 风险等级与业务含义
翻译成谁的数据、什么操作、停机还是泄露、是否不可逆。 |
| ☐ | 修复建议
按可执行性写：裁工具、加审批、检索过滤、拆分身份。不把“加强提示词”当成唯一措施。 |
| ☐ | 复测结果
对已修复项用同一条路径再走一遍。没复测的修复，只能算计划。 |
| ☐ | 管理层摘要
两页内说清：测了什么、最重要的三件事、建议的上线条件。 |

不应当被算作交付

- | | |
|--------------------------|---------------------|
| ☐ | 只有模型厂商的安全白皮书。 |
| ☐ | 只有扫描器对底层主机的发现。 |
| ☐ | 只有无法回放的现场演示。 |
| ☐ | 只有口头承诺“生产里会关掉这个工具”。 |

NEXT STEP

想验证你的 Agent 是否能够通过这些检查？

带上一个即将上线的 Agent、它的工具表和一张边界草图。我们用三十分钟把明显的阻塞项收口，并判断是否需要一次带证据包的正式评估。这不是产品演示，也不是口头保证。

预约一次 30 分钟 AI Agent 安全架构 Review

contact@yofunesec.com <https://yofunesec.com/ai-agent-security/>

评估落地页 · 申请评估或预约 Review